Are we there yet? Encoder-decoder neural networks as cognitive models of English past tense inflection

Are we there yet? Encoder-decoder neural networks as cognitive models of English past tense inflection

Maria Corkery mcorkery@inf.ed.ac.uk
Yevgen Matusevych yevgen.matusevych@ed.ac.uk
Sharon Goldwater sgwater@inf.ed.ac.uk

School of Informatics
University of Edinburgh

Abstract

The cognitive mechanisms needed to account for the English past tense have long been a subject of debate in linguistics and cognitive science. Neural network models were proposed early on, but were shown to have clear flaws. Recently, however, Kirov and Cotterell (2018) showed that modern encoder-decoder (ED) models overcome many of these flaws. They also presented evidence that ED models demonstrate humanlike performance in a nonce-word task. Here, we look more closely at the behaviour of their model in this task. We find that (1) the model exhibits instability across multiple simulations in terms of its correlation with human data, and (2) even when results are aggregated across simulations (treating each simulation as an individual human participant), the fit to the human data is not strong—worse than an older rule-based model. These findings hold up through several alternative training regimes and evaluation measures. Although other neural architectures might do better, we conclude that there is still insufficient evidence to claim that neural nets are a good cognitive model for this task.

1 Introduction

For over 30 years, the English past tense has served as both inspiration and testbed for models of language acquisition and processing (Rumelhart and McClelland, 1986; Pinker and Prince, 1988; Marcus, 1995; Plunkett and Juola, 1999; Pinker and Ullman, 2002; Albright and Hayes, 2003; Seidenberg and Plaut, 2014; Kirov and Cotterell, 2018; Blything et al., 2018, etc.). One of the most well-known debates centres on whether the apparently rule-governed regular past tense is indeed represented cognitively using explicit rules. Rumelhart and McClelland (1986) famously argued against this hypothesis, presenting a neural network model intended to capture both regular and irregular verbs with no explicit rules. However, Pinker and Prince (1988) presented a scathing rebuttal, pointing out both theoretical and empirical failures of the model. In their alternative (dual-route) view, the regular past tense is categorical and captured via explicit rules, while irregular past tenses are memorized and can (occasionally) generalize via gradient analogical processes (Pinker and Prince, 1988; Prasada and Pinker, 1993). Their arguments were so influential that although neural networks gained considerable traction in cognitive science more generally (Bechtel and Abrahamsen, 1991; McCloskey, 1991; Elman et al., 1996), many linguists dismissed the whole approach.[1]

With the recent success of deep learning in NLP, however, there has been renewed interest in exploring the extent to which neural networks capture human behaviour in psycholinguistic tasks (e.g., Linzen and Leonard, 2018; Linzen, 2019). In particular, Kirov and Cotterell (2018; henceforth K&C) revisited the past tense debate and showed that modern sequence-based encoder-decoder (ED) models overcome many of the criticisms levelled at Rumelhart and McClelland's model. Specifically, these models permit variable-length input and output that represent sequential ordering; can reach near-perfect accuracy on both regular and irregular verbs seen in training; and (using multi-task learning) can effectively generalize phonological rules across different inflections.

These primary claims are undoubtedly correct (and indeed, we replicate the accuracy results below). However, we take issue with another part of K&C's work, in which they claim that their ED model also effectively models human behaviour in a nonce-word experiment (i.e., *wug* test, described below). We explore the model's behaviour on this

[1] Though see Seidenberg and Plaut (2014), who argue that some of the core ideas, such as the focus on statistical learning, have nevertheless permeated the study of language.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3868–3877
Florence, Italy, July 28 - August 2, 2019.

task in detail, and conclude that its ability to model humans is considerably weaker than K&C suggest.

In particular, we begin by showing that multiple simulations of the same model (with different random initializations) result in very different correlations with the human data. To ensure that this instability is not just due to the evaluation measure, we introduce an alternative measure, but still find unstable results. We then consider whether treating individual simulations as individual participants (rather than as a model of the average participant) captures the human data better. This aggregate model does show some high-level similarities to the human participants: both model and humans tend to produce irregulars more frequently for nonce words that are similar to many real irregular verbs. However, the model is still poor at capturing fine-grained distinctions at the level of individual verbs. We conclude that, although deep learning approaches overcome many of the problems of earlier neural network models, there is still insufficient evidence to claim that they are good models of human morphological processing.

2 Background

2.1 Nonce word experimental data

Like K&C, we use data from two experiments run by Albright and Hayes (2003; henceforth A&H). In Experiment 1, using a dialogue-based prompt, A&H presented participants auditorily with nonce "verbs" that are phonotactically legal in English (e.g., *spling, dize*), and prompted participants to produce past tense forms of these verbs, resulting in a data set of **production probabilities** of various past tense forms. In Experiment 2, participants first produced each past tense form (as in Experiment 1) and were then asked to rate the acceptability of either two or three possible past tense forms for that verb—one regular, and one or two potential irregulars. For example, for *scride* /skrˈaɪd/, participants rated *scrided* /skrˈaɪdəd/ (regular), *scrode* /skrˈoʊd/ and *scrid* /skrˈɪd/ (irregular). This gives a data set of past tense form **ratings**.

Most of A&H's own analyses rely on the ratings data, but the ED model is a model of production, so we follow K&C and use the data from Experiment 1. The data is coded using the same set of suggested forms that were rated in Experiment 2: for each nonce word, A&H counted how many participants produced the regular form, the irregular form (or each of the two forms, if there are two), and "other" (any other past tense form that was not among those rated in Experiment 2). The counts are normalized to compute production probabilities for each output form.

The nonce words used by A&H were carefully chosen according to several criteria. First, they are phonologically "bland": i.e., not unusual-sounding as English words (as confirmed by a pre-test with participants). Second, as explained in the following section, they fall into several categories designed to test A&H's hypothesis that (contra Prasada and Pinker, 1993), *both* regular and irregular past tense forms exhibit gradient (and not categorical) effects.

2.2 A&H's model and islands of reliability

To explain the categories of nonce words (which we will refer to in our analyses), we briefly describe A&H's theory of past tense formation, which they implement as a computational model. The model postulates that speakers maintain a set of explicit structured rules that capture inflectional changes at different levels of generality. For example, a speaker might have rules such as:

- /∅/ → /əd/ if verb matches [X {/d/, /t/} ___] based on, e.g., *want, need, start*.
- /i/ → /ɛ/ if verb matches [X {/r/, /l/} ___ /d/] based on, e.g., *read, lead, breed*.

where X represents arbitrary phonological material and ___ is the location of the changing material. Each rule is given a confidence score based on its precision and statistical strength (the number of cases to which it could potentially apply). When a nonce word is presented, several rules may apply (e.g., the two rules above for *gleed*), and the goodness of each possible past tense is determined by the confidence score of the corresponding rule.

Crucially, A&H's model can learn multiple rules that all produce regular past tense forms, but with phonological contexts of different specificity, hence different confidence scores. Therefore, some nonce words may reside in so-called "islands of reliability" (IOR) for regular verbs: that is, there is an applicable regular rule that has a very high confidence score. Meanwhile other nonce words might also be considered regular, but with lower confidence. Thus, the model predicts gradient effects even for regular inflection. It also predicts gradient effects for irregular inflection, since there can be IORs for irregular rules as well.

To test these predictions, A&H chose four types of nonce words: those residing in an IOR for regu-

lars, for both regulars and irregulars, for irregulars only, or for neither. They also included several nonce verbs similar to *burn–burnt, spell–spelt*, and some that might potentially elicit single-form analogies. Their results (discussed further in Section 4) showed that the different IOR categories were indeed treated differently by participants.

2.3 Evaluating models

To go beyond coarse-grained analysis based on the IOR categories, both A&H and K&C evaluate their models by correlating model output with the human data at the level of individual past tense forms. Correlations are computed between the human data (either production probabilities or ratings) and the model scores for each form. The regulars and irregulars are treated separately. That is, the irregular correlation value is computed by considering the average human production probability (or rating) for each suggested irregular past tense, and comparing these with the model scores for those same forms. The correlation for regulars is computed analogously. Regulars and irregulars are treated separately because the scores for regulars are nearly always larger, so if all forms were considered at once, a baseline that simply assigned (say) 1 to regulars and 0 to irregulars would already achieve a high correlation with humans.

We initially follow K&C in computing the Spearman (rank) correlation against the production probabilities, and later also examine Pearson (linear) correlations and ratings data.

3 Methods

3.1 Model and hyperparameters

We adopt the encoder-decoder architecture used by K&C, as well as their implementation framework and hyperparameters. Encoder-decoder models are a type of recurrent neural network (RNN) introduced for machine translation (Sutskever et al., 2014) but also often used for other sequence-to-sequence transductions, such as morphological inflection and lemmatization (Kann and Schütze, 2016; Bergmanis and Goldwater, 2018). The encoder is an RNN that reads in the input sequence (here, a sequence of characters representing the phonemes in the present tense verb form) and creates a fixed-size vector representation of it. The decoder is another RNN that takes this vector as input and decodes it sequentially, outputting one symbol at each timestep (here, the phonemes of the past tense form). The ED model with attention (Bahdanau et al., 2015) is implemented in OpenNMT (Klein et al., 2017).[2] It has two bidirectional LSTM encoder layers and two LSTM decoder layers, 300-dimensional character embeddings in the encoder, and 100-dimensional hidden layers in the encoder and decoder. The Adadelta optimizer (Zeiler, 2012) is used for training, with the default beam size of 12 for decoding. The batch size is 20, and dropout is applied between layers with a probability of 0.3. Except where otherwise noted below, all models were trained for 100 epochs.

3.2 Training data

To compare our results to both A&H and K&C, we use their corresponding training sets, both based on data from CELEX (Baayen et al., 1995). A&H's training data contains all verbs listed in CELEX with a lemma frequency of 10 or more (4253 verbs, 218 of which are irregular). We use A&H's American English IPA phonemic transcriptions, to match the nonce word experiment (which was carried out with American English speakers), and also follow them in using the nonce words as the unseen test set rather than creating dev/test splits from the CELEX data. As argued by A&H, adult English speakers will have been exposed to all of the real verbs many times and would be able to correctly produce the past tense of all of them. Adults' generalization to nonce words is therefore predicated on their knowledge of this entire training set (including, crucially, all of the irregular forms).

For our second training set, we obtained the data from K&C, which is a subset of A&H's: it contains 4039 verbs, 168 of which are irregular—that is, 50 real irregular verbs are missing. Examples of verbs that are missing from the K&C data include *do–did* and *use–used*. K&C also randomly divided their data into training, development, and test sets, but we weren't able to obtain these splits, so (since we are using the nonce words for test data) we simply use all 4039 verbs as training data. We include results using the K&C's data mainly to allow closer (though still not exact) comparison with their work, but we feel that A&H's training data, which includes all the irregulars, more accurately reflects adult linguistic exposure.

It has been argued that morphological generalization in humans is governed by type frequencies

[2] In early tests, we also tried the Nematus toolkit with hyperparameters following (Kann and Schütze, 2016; Bergmanis and Goldwater, 2018); the pattern of results was similar.

Rank	nold /nˈoʊld/	Probability
1	**nolded** /nˈoʊldəd/	0.9869
2	nelt /nˈɛlt/	0.0120
3	neelded /nˈiːldəd/	0.0004
4	nelded /nˈɛldəd/	0.0004
5	**neld** /nˈɛld/	0.0001
Rank	murn /mˈərn/	Probability
1	**murned** /mˈərnd/	0.8636
2	**murnt** /mˈərnt/	0.1363
3	murn /mˈərn/	<0.0001
4	murnaid /mˈərneɪd/	<0.0001
5	murnoo /mˈərnuː/	<0.0001

Table 1: Top 5 outputs from two sample beams, for the nonce words *nold* and *murn*. Past tenses suggested by A&H are bolded. For *nold*, one suggested past tense form, *nold* /nˈoʊld/, is missing from the top 5.

rather than token frequencies (Bybee and Thompson, 1997; Pierrehumbert, 2001). Modelling evidence, including from A&H, also supports the idea that token frequencies are ignored or severely downweighted (i.e., effectively using log frequencies: O'Donnell, 2015; Goldwater et al., 2006). We therefore follow A&H and K&C in training our models on the list of distinct word types, with each type occurring once in the training data.

3.3 Evaluation

We report three different evaluation measures. First, we compute **training set accuracy**: the percentage of verbs in the training data for which the model's top-ranked output is the correct past tense form. This is largely a sanity check and test of convergence: a fully-trained model of adult performance should have near-perfect training set accuracy.

Next, as described in Section 2.3, we report Spearman's rank **correlation** (ρ) of the model's probabilities for the various nonce past tense forms with the human production probabilities. The probability for each suggested past tense form was obtained by forcing the model to output that form (e.g., providing *scride* as input and forcing it to output *scrid*). This made it possible to get probabilities for forms that did not occur in the beam (the list of most likely forms output by the model).

Finally, we introduce a third measure, motivated further in Section 4.1, **complete recall@5**:

$$\text{CR@5} = \frac{1}{n} \times \sum_{i=1}^{n} [S_i \subseteq B_i] \qquad (1)$$

where n is the total number of nonce verbs, S_i is the set of A&H's suggested past tense forms for verb i, B_i is the set of the top five verbs in the model's beam for i, and $[S_i \subseteq B_i] = 1$ if all verbs from S_i appear in B_i, and 0 otherwise. For example, a model which only processed the two verbs in Table 1 would have a CR@5 of 0.5, since the beam includes all suggested past tenses for *murn* (*murned, murnt*), but not for *nold* (*nolded, nold, neld*).[3]

Data	all	regular	irregular
K&C	99.79 (0.05)	99.92 (0.04)	96.90 (1.06)
A&H	99.51 (0.04)	99.86 (0.07)	92.98 (1.18)

Table 2: Mean training set accuracy (in %, with standard deviations in brackets), averaged over 10 runs for each training set with different random seeds. Oracle accuracy is 99.85% on the K&C data and 99.55% on the A&H data, due to homophones and forms with multiple past tenses. In order to do better on irregulars, the model would have to get more of the regulars wrong.

4 Experiments

4.1 Experiment 1: Model variability

Our first experiment aims to replicate K&C's results showing that (a) the model is able to produce the past tense forms of training verbs with near-perfect accuracy, and (b) its correlation with human data on the nonce verb test set is higher than that of A&H's model. In K&C's paper, these results were based on a single trained model. Here we trained 20 models (10 on each training set) initialized with different random seeds.

Accuracy Table 2 lists the mean and standard deviation of training set accuracy for each of the two training sets. It is not possible to get 100% accuracy because the training sets contain some homophones with different past tenses (e.g., *write–wrote* and *right–righted*), and some verbs which have two possible past tenses (e.g., *spring–sprung* and *spring–sprang*). Nevertheless, the models get very close to the best possible accuracy, confirming K&C's finding that they learn both regular and irregular past tenses of previously seen words within 100 epochs. Example convergence plots are shown in

[3] Not all of A&H's suggested forms were actually produced by participants, but all of them seem plausible and we felt that a good model should rank them higher than most other potential past tenses, i.e., they should be included within a small beam size. Indeed, in cases where they are not (e.g., *nold* in Table 1) we do typically see much less plausible forms (such as *neelded*) included in the beam.

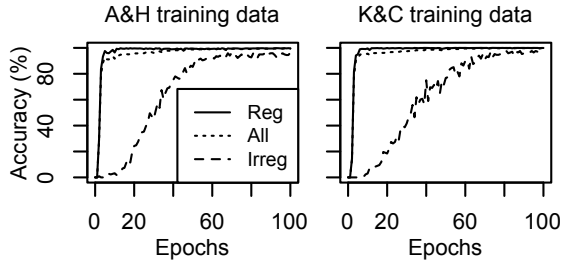


Figure 1: Accuracy values on the training set during training for one model per training set.

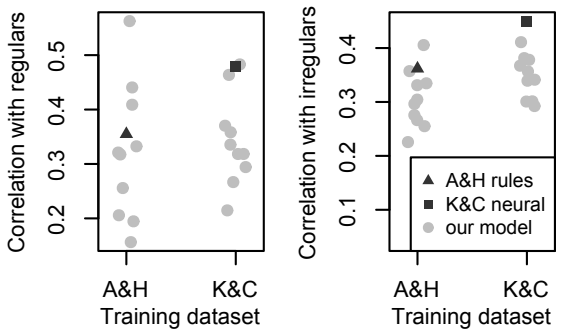


Figure 2: Spearman correlation coefficients between model scores and human production probabilities, using the A&H and K&C training data. Values reported by K&C and A&H are shown in addition to those of our models. Horizontal jitter is added for readability.

Figure 1, illustrating that the models learn regular verbs very quickly, and irregular verbs more slowly, but both are learned well after 60–80 epochs.

Correlation Despite having consistently high accuracy on real words, Figure 2 shows that models with different random initializations vary considerably in their correlation with human speakers' production probabilities on nonce words, from 0.15 to 0.56 for regulars, and from 0.23 to 0.41 for irregulars. K&C's reported results are at the high end of what we obtained, suggesting that they are likely not representative.

On the other hand, we were concerned that the variability in the correlation measure might be due to an artefact: the vast majority of the beams returned by the model assign very high probability ($> 98\%$) to the top item and little mass to anything else (as in the first example in Table 1).[4] Since the correlation measure is computed across different nonce forms, tiny changes in the beam probabilities of one nonce verb could change the ranking of (say) its regular past with respect to the regular past of another nonce word, even if the relative ranking of forms within each nonce's beam stayed the same.

[4]The skewedness of the beams is likely because of the training/testing scenario, where the model is effectively asked to do different tasks: at training time, it is trained to produce one correct past tense, while at test time, it's expected to produce a probability distribution over potential nonce past tenses. We could surely produce better matches to the human probability distributions by training directly to do so, but that wouldn't make sense as a cognitive model, since human learners are exposed only to correct past tenses, not to distributions.

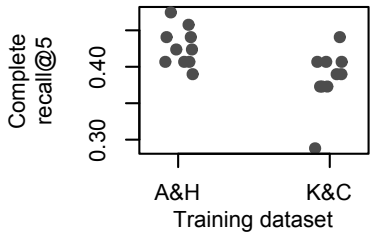


Figure 3: Complete recall@5 for 20 models with different random seeds (10 with each training dataset). Horizontal jitter is added for readability.

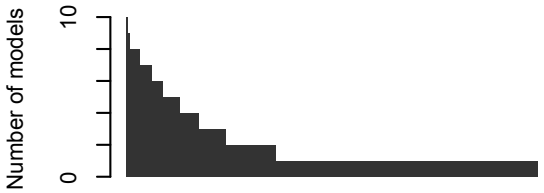


Figure 4: The number of models (of the 10 trained on the A&H dataset) which agree on the second-place past tense form. The X-axis shows 281 different past tense forms (for 59 nonce words in the present tense), and the Y-axis shows, for each form, how many times a model places it in the second position in the beam.

CR@5 and second best forms The above observation motivated the CR@5 measure (Section 3.3). Rather than measuring the relative probabilities of past forms across different verbs, CR@5 considers the relative rankings of different past forms for each verb. However, CR@5 also yielded unstable results: 39–47% on A&H's data, and 29–44% on K&C's data, as shown in Figure 3.

As a final exploration of the models' instability across different simulations, we looked at how often the models agree with each other on the verb occupying the first and the second position in the beam. While there is very high agreement on the most likely form (top of the beam) across the simulations—usually a regular past tense—very few forms in the second position are the same across simulations (see Figure 4).

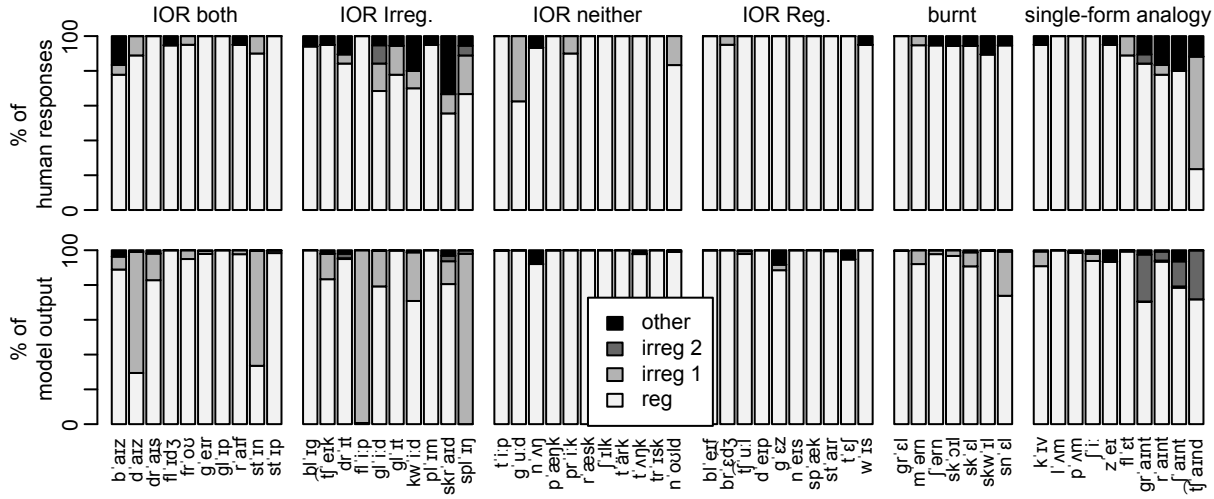


Figure 5: Percentage of regular, irregular, and "other" responses produced by humans (top) and the model (bottom). Each of the six blocks corresponds to a different category of nonce words (see Section 2.2).

Summary To recap, we find similar training set accuracy to what K&C reported, but the correlation scores between the model and the human data are generally lower, and the model exhibits unstable behaviour across different simulations. However, the unstable behaviour can potentially be accounted for, if each simulation is interpreted as an individual participant rather than as a model of the average behaviour of all participants. In that case, we should aggregate results from multiple simulations in order to compare them to the human results, since production probabilities from A&H's experiment were obtained by aggregating data over multiple participants. The next experiment examines this alternative interpretation.

4.2 Experiment 2: Aggregate model

To simulate A&H's production experiment with each simulation as one participant, we trained 50 individual models on the A&H training data[5] using the same architecture and hyperparameters as before. We then sampled 100 past tense forms for each verb from each model's output probability distribution. Each of the 5000 output forms (100 each from 50 simulated participants) was categorized either as (a) the verb's regular past tense form, (b–c) the first or second irregular past tense form suggested by A&H, or (d) any other possible form.

For the aggregate model, the correlation measure is the only evaluation that makes sense. For regulars, correlation with the human production probabilities was higher than in the previous experiment (0.45 vs. an average of 0.28 in Experiment 1), but for irregulars it was lower (0.19 vs. 0.22 in Experiment 1). The differences between the humans and aggregate model are clear from Figure 5, which shows the distribution of various past tense forms for both model and humans. For example, in only one case did the humans produce an irregular more frequently than the regular (no-change past *chind* for present *chind*), whereas there are several cases where the aggregated model does so. Moreover, for the word *chind* itself, the model prefers *chound* rather than *chind*.

In the previous experiment, we saw that individual models often rank implausible past tenses higher than plausible ones. However, we see here that on aggregate nearly all the model's proposed past tenses are those suggested by A&H. Apparently, the unstable beam rankings wash out the implausible forms, i.e., the plausible forms on average occur nearer the top of the beam than any particular implausible form. In fact, the model actually produces fewer "other" forms than the humans.

We also looked at the model's average production of regular and suggested irregular forms for each of the six categories in Figure 5. The results, shown in Figure 6, indicate that the model does capture the main trends seen in humans across these categories, but overall it is more likely to produce irregular forms. Together with the low overall correlation to human results and obvious differences at the fine-grained level, these results suggest that there are serious weaknesses in the ED model, even when results are aggregated across simulations.

[5] In the absence of clear differences between the model's performance on A&H's vs. K&C's data in Experiment 1, we only use the more complete A&H dataset henceforth.

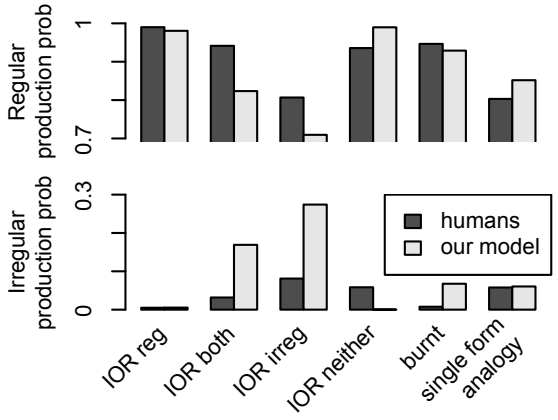


Figure 6: Mean production probabilities for regulars (top) and A&H's suggested irregulars (bottom) in each of A&H's categories of nonce words, for humans and for the aggregated ED model.

Data	Cor.	Verbs	A&H	Individ.	Agg.
Production	ρ	reg.	.35	.32 (.12)	**.45**
		irreg.	**.36**	.31 (.05)	.19
	r	reg.	**.62**	.16 (.09)	.30
		irreg.	.14	.16 (.03)	**.17**
Rating	ρ	reg.	**.55**	.32 (.09)	.43
		irreg.	**.57**	.39 (.08)	.31
	r	reg.	**.71**	.34 (.07)	.40
		irreg.	**.48**	.35 (.06)	.40

Table 3: Correlations (using Spearman's ρ and Pearson's r) between the models' output probabilities vs. human production probabilities and rating data. The data for the individual model is an average over 10 simulations (standard deviation shown in brackets). Highest correlation in each line is shown in bold.

5 Further analyses

5.1 Is the model overfitting?

We began by assuming that models should be trained at least until they achieve perfect performance on the training set, but perhaps 100 epochs is too much, and the model is just overfitting. Training for less time might produce less skewed beam probabilities, more stable beam rankings, and perhaps better correlations with the human data.

To investigate this possibility, we took the 10 models originally trained on the A&H dataset and computed the correlation with human data for regulars and irregulars after every 10 epochs of training. The highest correlation is achieved after only 10 epochs (0.47 for regulars and 0.50 for irregulars) and the beam probabilities are indeed less skewed: the average probability of the top ranked output is 0.92 after 10 epochs, vs. 0.97 after 100 epochs. However, the models average only 6.5% accuracy on the real irregular words after 10 epochs, so it is difficult to argue that these are good models of human behaviour.[6] It seems that the ED model displays a fundamental tension between correctly modelling humans on real words and nonce words.

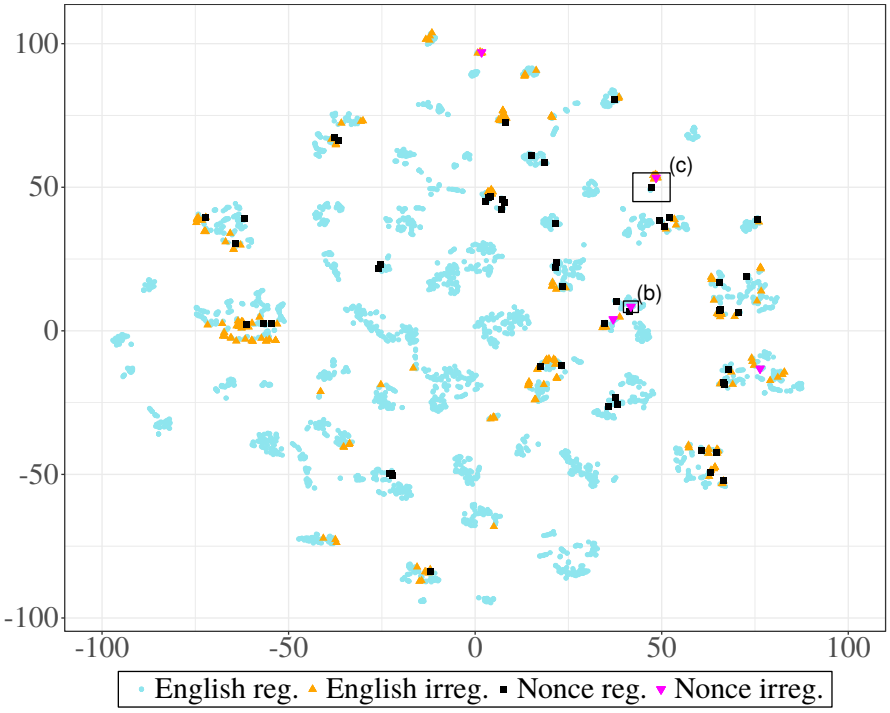
5.2 Rating data and correlations

We have so far evaluated all models against human production data. However, the A&H model outputs unnormalized scores, so arguably it makes more sense as a model of ratings. A&H also originally evaluated it using Pearson correlation. For completeness we report in Table 3 the correlations for all models on both ratings data and production data, using both Spearman and Pearson coefficients. We find that the A&H model does score better against ratings data, although surprisingly the ED models do too. More importantly, though, the A&H model fits the human data best on 6 out of 8 measures.

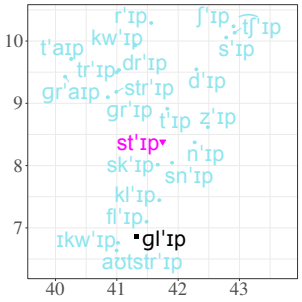
5.3 What is the model learning?

To examine the representations acquired by the model, we extract vectors from the encoder's hidden state. As the encoder is a bidirectional LSTM, we concatenate the two states at the last time step (after training on the A&H data). Figure 7a shows a t-SNE visualization of hidden state vectors for both real and nonce verbs in one of our simulations. The model clearly groups the verbs into small clusters, and Figures 7b–c show that this clustering is based on the verbs' trailing phonemes, including some structure withing the clusters: e.g., *strip* /strˈɪp/, *grip* /grˈɪp/, and *trip* /trˈɪp/ are next to each other in Figure 7b, and so are *clip* /klˈɪp/, *flip* /flˈɪp/, and *glip* /glˈɪp/. It is not so clear, however, how the model decides on whether to produce a regular or an irregular form for nonce verbs. We do see some evidence in Figure 7c that nonce verbs similar to regular English verbs yield a regular form (note the regular neighbours of *nung* /nˈʌŋ/), and the same holds for irregulars (note the irregular forms around *spling* /splˈɪŋ/, for which the model produced *splung*). However, the model also produces an irregular form (*stup* /stˈʌp/) for *stip* /stˈɪp/, which is clearly surrounded by regular En-

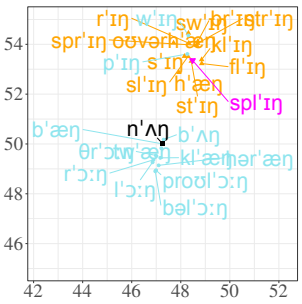
[6] Early exposure to more irregulars could help in principle, so we also tried training the models on token or log token frequencies rather than type frequencies, but the resulting models' correlations with production probabilities were no higher than models trained on type frequencies (the same for log tokens, and lower for tokens).



(a) All verbs.



(b) Zooming in on /st'ɪp/.



(c) Zooming in on /n'ʌŋ/.

Figure 7: A t-SNE plot of encoder state vectors for regular and irregular verb forms. (a) shows an an overview of all (real and nonce) verbs, and (b) and (c) zoom in on the boxed areas in (a).

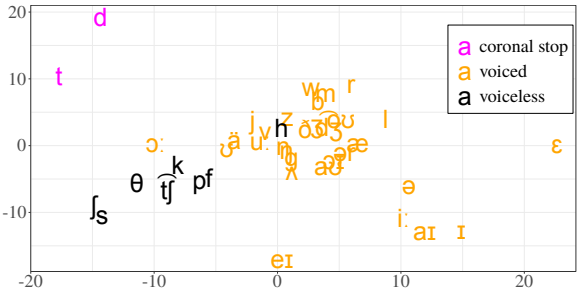


Figure 8: PCA plot of character-level (phoneme) vectors extracted from the decoder's hidden state. The phonemes are coloured based on the three different regular past-tense suffixes they would be followed by.

glish verbs in Figure 7b.

We also tested whether the clustering by trailing phonemes is simply an artefact, by training another model on data where we reversed the order of the input phonemes in all cases (e.g., /w'ɪʃ/–/w'ɪʃt/ [*wish–wished*] becomes /ʃɪ'w/–/tʃɪ'w/). This time, verbs were grouped based on their *leading* phonemes—that is, the endings of the original verbs—suggesting that the model finds the regularities in the data regardless of the order of phonemes.

Finally, we investigated the model's phoneme representations, expecting a clustering corresponding to the three types of phonemes that trigger different endings in regular past tense forms: /-ɪd/ after coronal stops /t/ and /d/, /-d/ after voiced consonants and vowels, and /-t/ after voiceless consonants. We extract character-level vectors from the decoder hidden state, apply PCA (which worked better than t-SNE in this case) and visualize the resulting vectors. Figure 8 shows that the expected pattern has emerged (except for /h/ in the 'voiced' cluster, but this phoneme never appears at the end of English words).

6 General discussion and conclusions

Our results confirm that, unlike earlier neural net models, the ED model has no trouble learning the past tense forms of verbs it is trained on. We found, however, that its behaviour on nonce verbs does not correlate with human experimental data as well as K&C's results implied, and indeed not as well as that of A&H's much earlier rule-based model.

One issue in particular seems to be overproduction of irregulars, which the model consistently prefers to regulars for four verbs (7% of considered nonce verbs), while humans nearly always prefer the regular form. This was an issue with earlier neural net models as well (Plunkett and Juola, 1999). On the other hand, when the model

outputs something other than the regular form, its choices are plausible. This was not true for earlier models: Plunkett and Juola's model often chose the wrong regular suffix (with incorrect voicing in the final phoneme), and Rumelhart and McClelland's (1986) model failed to produce regular endings for nonce verbs (Prasada and Pinker, 1993; Marcus, 1998). Here, we see from both our model's output and its internal representations that it has correctly identified the necessary voicing distinctions and that nonce words trigger similar representations and behaviour to real words. In future, a stricter test might use nonce words that are intentionally less similar to real words (e.g., the example from Prasada and Pinker (1993): *to out-Gorbachev*).

It is also worth pointing out that the ED model, unlike A&H's model and many earlier connectionist models, is fed raw phonemes (rather than the phonemes' distinctive features) as input. Although it learns some of the relevant features anyway, it would be interesting to see whether its behaviour becomes more human-like if the correct features are provided in the input.

Although our paper has revealed a number of weaknesses of the ED model, we do agree with K&C that neural network-based cognitive models of inflection deserve re-evaluation in light of recent technical advances. There are many other potential architectures and modelling decisions that could be explored, as well as other behavioural data such as developmental patterns (Blything et al., 2018; Ambridge, 2010) and inflection in other languages (e.g., Clahsen et al., 1992; Ernestus and Baayen, 2004). As noted by Seidenberg and Plaut (2014), models' failures as well as successes can be informative, and we hope that our detailed exploration of the ED model's behaviour will inspire future developments in these models, both for cognitive modelling and NLP.

Acknowledgements

This work was supported in part by a James S McDonnell Foundation Scholar Award (220020374).

References

Adam Albright and Bruce Hayes. 2003. Rules vs. analogy in English past tenses: a computational/experimental study. *Cognition*, 90:119–161.

Ben Ambridge. 2010. Children's judgments of regular and irregular novel past-tense forms: New data on the English past-tense debate. *Developmental Psychology*, 46:1497–1504.

R. Harald Baayen, Richard Piepenbrock, and Leon Gulikers. 1995. CELEX2 LDC96L14. Web Download. Linguistic Data Consortium, Philadelphia, PA.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*, San Diego, CA, USA.

William Bechtel and Adele Abrahamsen. 1991. *Connectionism and the mind: An introduction to parallel processing in networks*. Basil Blackwell, Oxford, England.

Toms Bergmanis and Sharon Goldwater. 2018. Context sensitive neural lemmatization with Lematus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1391–1400. Association for Computational Linguistics.

Ryan P. Blything, Ben Ambridge, and Elena V.M. Lieven. 2018. Children's acquisition of the english past-tense: Evidence for a single-route account from novel verb production data. *Cognitive Science*, 42:621–639.

Joan Bybee and Sandra Thompson. 1997. Three frequency effects in syntax. In *Proceedings of the 23rd Annual Meeting of the Berkeley Linguistics Society*, pages 378–388. Berkeley Linguistics Society, Berkeley, CA.

Harald Clahsen, Monika Rothweiler, Andreas Woest, and Gary F. Marcus. 1992. Regular and irregular inflection in the acquisition of German noun plurals. *Cognition*, 45:225–255.

Jeffrey Elman, Elizabeth Bates, Mark H. Johnson, Anette Karmiloff-Smith, Domenico Parisi, and Kim Plunkett. 1996. *Rethinking innateness: A connectionist perspective on development*. MIT Press, Cambridge, MA.

Mirjam Ernestus and R. Harald Baayen. 2004. Analogical effects in regular past tense production in Dutch. *Linguistics*, 42:873–903.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2006. Interpolating between types and tokens by estimating power-law generators. In *Advances in NIPS-18*, pages 459–466. Curran Associates, Inc., Red Hook, NY.

Katharina Kann and Hinrich Schütze. 2016. MED: The LMU system for the SIGMORPHON 2016 shared task on morphological reinflection. In *Proceedings of the 14th Annual SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 62–70. Association for Computational Linguistics, Stroudsburg, PA.

Christo Kirov and Ryan Cotterell. 2018. Recurrent neural networks in linguistic theory: Revisiting Pinker and Prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, 6:651–665.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, System Demonstrations*, pages 67–72. Association for Computational Linguistics.

Tal Linzen. 2019. What can linguistics and deep learning contribute to each other? Response to Pater. *Language*, 95:99–108.

Tal Linzen and Brian Leonard. 2018. Distinct patterns of syntactic agreement errors in recurrent networks and humans. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, pages 692–697. Cognitive Science Society, Austin, TX.

Gary F. Marcus. 1995. The acquisition of the English past tense in children and multilayered connectionist networks. *Cognition*, 56:271–279.

Gary F. Marcus. 1998. Can connectionism save constructivism? *Cognition*, 66:153–182.

Michael McCloskey. 1991. Networks and theories: The place of connectionism in cognitive science. *Psychological Science*, 2:387–395.

Timothy J. O'Donnell. 2015. *Productivity and reuse in language: A theory of linguistic computation and storage*. MIT Press, Cambridge, MA.

Janet Pierrehumbert. 2001. Stochastic phonology. *Glot International*, 5:195–207.

Steven Pinker and Alan Prince. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28:73–193.

Steven Pinker and Michael T. Ullman. 2002. The past and future of the past tense. *Trends in Cognitive Sciences*, 6:456–463.

Kim Plunkett and Patrick Juola. 1999. A connectionist model of English past tense and plural morphology. *Cognitive Science*, 23:463–490.

Sandeep Prasada and Steven Pinker. 1993. Generalisation of regular and irregular morphological patterns. *Language and Cognitive Processes*, 8:1–56.

David E. Rumelhart and James L. McClelland. 1986. On learning the past tenses of English verbs. In James L. McClelland, David E. Rumelhart, and the PDP Research Group, editors, *Parallel distributed processing: Explorations in the microstructure of cognition*, chapter 18, pages 216–271. MIT Press, Cambridge, MA.

Mark S. Seidenberg and David C. Plaut. 2014. Quasiregularity and its discontents: The legacy of the past tense debate. *Cognitive Science*, 38:1190–1228.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, pages 3104–3112. Curran Associates, Inc., Red Hook, NY.

Matthew D. Zeiler. 2012. ADADELTA: An adaptive learning rate method. *Computing Research Repository*, arXiv:1212.5701.